Attention Please: Deep Dive into Image Captioning Model Efficiency
Mithun “Mit” Ramesh

This project investigated the performance of transformer-based models in image captioning tasks, critically examining the limitations of generalized models like VGG19, ResNet, and InceptionX. We created a Transformer seq2seq model from scratch in order to replicate our own image caption prediction model that was trained on the Flickr8K Dataset. In creating this self-created model, we experimented with many hyperparameters and experimental decisions (Different inference algorithms, using RNN/LSTM, etc.). We then compared this performance to the performance of un-tuned pre-trained models with a fine-tuned model on specific datasets. Our objectives were to evaluate the state-of-the-art in image captioning and assess the adaptability of these models to specialized datasets, a crucial aspect for applications like medical imaging. This was a group project, and for my portion of the capstone project, I attempted to expand the original Flickr8k dataset to include the Flickr30k dataset and the COCO120k dataset; the preprocessing classes can be found in our github repository. To improve the training efficiency, I implemented multithreading on both the Flickr8k and Flickr30k datasets. Additionally, I developed a tokenizer and embedding class to process the textual data in the datasets for the from-scratch model. To evaluate the performance of our seq2seq model, I also implemented the Greedy Search method as an evaluation method.